Uncooperative Person Recognition Based on Stochastic Information Updates and Environment Estimators

Hye-Jin Kim, Dohyung Kim, Jaeyeon Lee, and Il-Kwon Jeong

We address the problem of uncooperative person recognition through continuous monitoring. Multiple modalities, such as face, height, clothes color, and voice, can be used when attempting to recognize a person. In general, not all modalities are available for a given frame; furthermore, only some modalities will be useful as some frames in a video sequence are of a quality that is too low to be able to recognize a person. We propose a method that makes use of stochastic information updates of temporal modalities and environment estimators to improve person recognition performance. The environment estimators provide information on whether a given modality is reliable enough to be used in a particular instance; such indicators mean that we can easily identify and eliminate meaningless data, thus increasing the overall efficiency of the method. Our proposed method was tested using movie clips acquired under an unconstrained environment that included a wide variation of scale and rotation; illumination changes; uncontrolled distances from a camera to users (varying from 0.5 m to 5 m); and natural views of the human body with various types of noise. In this real and challenging scenario, our proposed method resulted in an outstanding performance.

Keywords: Normalization, stochastic update, environment estimator, person recognition.

Manuscript received Feb. 1, 2014; revised Feb. 2, 2015; accepted Feb. 3, 2015.

This work was supported by the Cross-Ministry Giga KOREA Project of the Ministry of Science, ICT and Future Planning, Rep. of Korea (GK14C0100, Development of Interactive and Realistic Massive Giga-Content Technology).

Hye-Jin Kim (corresponding author, marisan@etri.re.kr), Dohyung Kim (dhkim008@etri.re.kr), Jaeyeon Lee (leejy@etri.re.kr), and Il-Kwon Jeong (jik@etri.re.kr) are with the SW·Content Research Laboratory, ETRI, Deajeon, Rep. of Korea.

I. Introduction

Person recognition is one of the most important issues in video and real-time applications. It is well known that on-line person recognition performs worse than off-line recognition. In the case of on-line recognition, though most previous works have made use of various modalities (audio and video) when attempting to identify a person, they have done so without regard for time asynchrony. Consequently, attempts to utilize a combination of audio and visual modalities have only led to a reduction in person recognition performance.

In the case of a real environment, uncooperative person recognition is a challenging issue owing to its many difficulties, such as pose and view variation and time asynchrony between different modalities. For example, it is hard to obtain a 100% face detection rate even when incorporating goal-related knowledge [1]. Difficulties, such as illumination, pose variation, time asynchrony, and human scale changes, exist that affect recognition steps; hence, attempts have been made to alleviate these problems by combining tracking and recognition.

Temporal modalities obtained by tracking an individual provide the key to improving the person recognition performance of the individual incrementally. In [2], human location using a radar sensor is fused with identification based on clothes color. Hybrid sensors are used in tracking and interactive recognition [3]. The authors in [4]–[5] described simultaneous face tracking and recognition. These works couple a recognizer and tracker and hence update both at the same time. These approaches accumulate information regardless of the quality of contents, which causes a long-term

failure in person recognition. Moreover, these approaches do not deal with the problem of time asynchrony.

In [7], the authors track and recognize faces with visual constraints in real-world video and achieve high recognition rates. In one of the video sequences used in [7], almost all views of faces are of a frontal nature; for example, people are sitting on a chair in front of the camera. However, in the real world, most people move without restriction; therefore, one cannot guarantee that a person's face will always be clearly visible. In addition, the sizes of the faces of the people used in [7] are comparatively larger than those in our "real-world-like environment" database and hence are much less varying. In short, the work in [7] does not deal with dramatic variations in face sizes and viewpoints.

In [8], a person is re-identified using chromatic contents of clothes and body structure information. Using an appearance-based approach, they achieve robustness to variations in pose, viewpoint, and illumination. This paper mostly deals with the whole body of a person so as to capture the overall clothes color. Under this approach, body information can be easily preserved because whole scenes always contain the entire body of the target person. As mentioned previously, however, people in daily life move around; thus, it is essential to deal with variations in a person's structural deformation.

There have been many researches that have investigated evidence accumulation [8]–[11] and cue integration [12]–[14]. However, these works are not direct studies on fusion schemes under continuous monitoring. Moreover, they do not deal with asynchronous data or with images where the quality of the modalities used within the images is not good enough; such images occur frequently in real time. To apply a person recognition system to a real-time application, it is necessary to find a method that can handle both the asynchrony of data and a lack of information related to modalities.

Another important issue is how to deal with a large number of redundant data of a video clip or streaming ones. In the case of image data, for example, we can easily obtain sixty images (frames) per second, but we don't necessarily need to use all of these images in our efforts to recognize a person — this is because a person cannot move far within 1/60 of a second; hence, most of the images will probably contain the same person. In addition, some of the images may contain meaningless information. Therefore, it is better to be able to choose data selectively; thus, to be able to make such a decision, it is necessary to contrive a new set of parameters.

When it comes to multi-modal fusion, it is worth knowing how to integrate differently-distributed evidence. In [15], Lee and others proposed a discrete PDF to compensate the distribution differences. This method shows a better performance than parametric approaches and can reduce the normalization error between independent modalities.

In this paper, we present a novel method to recognize an uncooperative person based on stochastic information updates and environment estimators using multiple modalities, such as face, height, clothes color, and voice. Our method includes density normalization, environment estimators, and stochastic information updates. The process of the proposed method is described in detail in Sections II and III.

II. Proposed Method

The proposed method considers unconstrained environments; that is, where people can move without restrictions, such as body pose and position in a space. Our method consists of density normalization, environment estimators, and a stochastic information update (SIU) algorithm. Density normalization provides us with the capability to compare different modalities. The environment estimators play a crucial role in dealing with the large, but often redundant, amounts of data. The estimators indicate when certain modalities are worthy of use in recognizing a person, regardless of data asynchrony. The use of continuous monitoring enables us to use an SIU algorithm, and the environment estimators provide the solution to the problem of time mismatches between the different modalities. Figure 1 illustrates an overview of the proposed algorithm. Using a training database of multiple modalities, we first trained recognizers and extracted normalization parameters. Then, to a given an input dataset (image and voice), we applied four kinds of recognizers while tracking a target person. After this recognition step, we simultaneously apply *Charf*-based

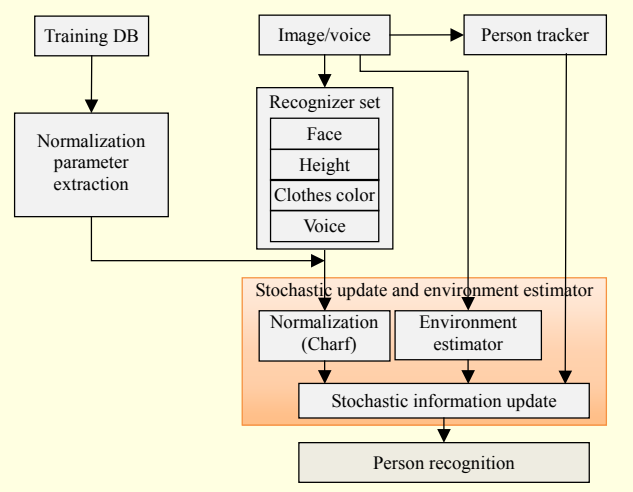


Fig. 1. Overall process of proposed approach. Inputs are the modalities. We achieve an increase in person recognition performance by using a combination of environment estimators and stochastic information updates, which can be implemented because of continuous monitoring.

ETRI Journal, Volume 37, Number 2, April 2015
http://dx.doi.org/10.4218/etrij.15.0114.0037

normalization and environment estimators to the input data. In addition to this, we use the SIU algorithm to continuously update the likelihood of each modality on a frame-by-frame basis so as to improve the likelihood of successfully recognizing the tracked person.

1. Normalization

Normalization methods for multi-modal person recognition make use of predominantly mean and variance–based techniques. This is because data distributions are conventionally assumed to be Gaussian. Most of them, however, have a rather skewed distribution because the recognizers used in their composition affect the data distributions, as shown in Fig. 2 and [15]. To reflect a real distribution, we adopt the *Charf* function [16] to normalize the different likelihoods of the modalities.

For a given modality, M_i, $i \geq 0$, $i \in \mathbb{Z}^+$, we denote its likelihood by m_i. In addition, we denote an input pattern by $\mathbf{x}$ and a class by ω_i, where a class corresponds to a target person. Mathematically, m_i should be between zero and one, but in real practice, its value is unbounded.

The main goal of [16] is to normalize the output of each modality such that it reflects the actual modal performance

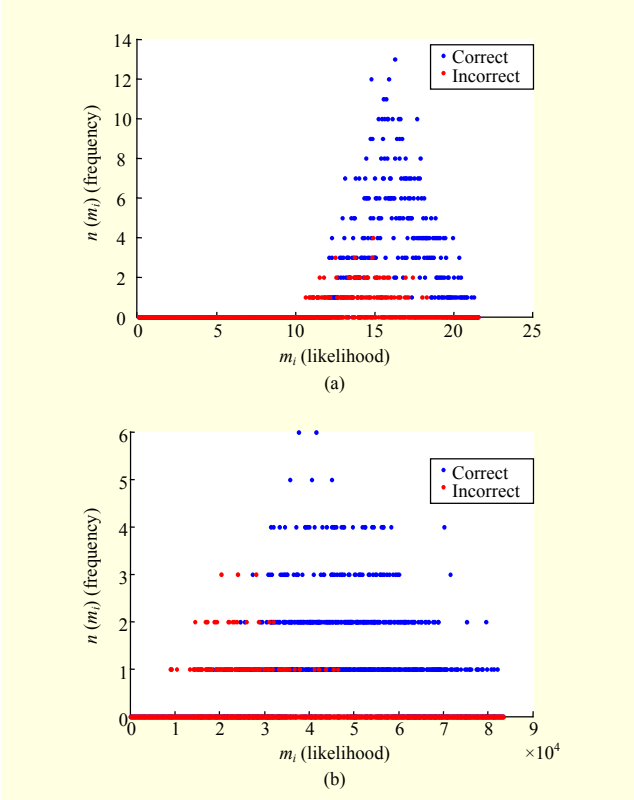


Fig. 2. Distributions of correctly and incorrectly recognized data: (a) speaker recognizer and (b) face recognizer.

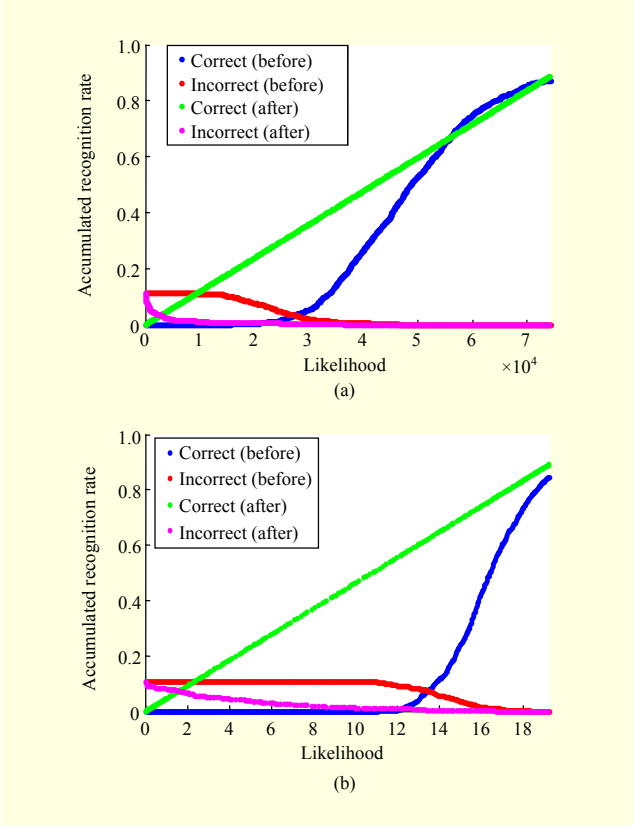


Fig. 3. Accumulated recognition rates after applying the Charf function. Correctly recognized outputs and incorrect ones are presented: (a) face recognition and (b) speaker recognition.

for each output value. That is, $m_i > m_j$ should imply $P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x})$. This normalization process is as follows:

1. Obtain the accumulated recognition rates for the training data.
2. Calculate the transfer function (Charf).

$$\text{Charf}(m_i) = m_{i\max} \times \frac{\sum_{j=0}^{i} n_{\text{correct}}(m_j)}{N} - m_i, \quad (1)$$

where N is the total number of patterns and $n_{\text{correct}}(m_j)$ is the number of correctly recognized trained patterns where $0 \leq j \leq i, j \in \mathbb{Z}^+$.

3. Normalize the data using the Charf function.

$$m_i^{\text{new}} = m_i + \text{Charf}(m_i). \quad (2)$$

From the Charf function, the normalization equation is derived. The independently distributed speaker and face likelihoods in Fig. 2 are different, but their respective accumulated recognition rate distributions are similar after applying the Charf function, as shown in Fig. 3.

2. Environment Estimators

The environment estimators determine whether a modality

is useful. Here, "environment" implies the conditions of a recognizer, such as time asynchrony and the distance between a user and a sensor.

The time asynchrony between audio and video makes it hard to apply a fusion method in real time. For example, video-based data, such as face, height, and clothes color, are generated at 30 frames per second, whereas speech data occurs every two to four seconds since speaker recognition performance is dependent upon utterance length [17].

The conditions for good accuracy of person recognition conflict between the modalities. For instance, if a person is close to the camera (such as closer than 3 m), then a face recognizer may perform better than a height recognizer, or vice versa. To overcome this kind of problem, we demonstrated in [18] the use of an environment estimator, which helped to select informative frames by estimating the quality of the face.

When it comes to person recognition under real circumstances, it is better to make use of as many environment estimators as possible. For example, in our study, we made use of environment estimators for the size of the face (E_F), the illumination changes (E_V), and the SNR ratio of the audio applied in (3)–(5) (E_A). These parameters are applied to the experiments in our study using the thresholds δ_F, δ_V, and δ_A, respectively. Under the assumption that the data used during the enrollment stage are in good condition, such data was compared with the test data. The equation for the audio environment estimator E_A is

$$E_A = -\log_{10} \frac{\sum \mathbf{x}_{ref}(t)^2}{\sum \mathbf{x}_{noise}(t)^2}, \tag{3}$$

where $\mathbf{x}_{ref}(t)$ and $\mathbf{x}_{noise}(t)$ are audio signals. After detecting the end-point detection of the signals, E_A was found to range from the beginning of a speech signal to its first peak (0.5 s before the first peak, for example). The registered voice is used as $\mathbf{x}_{ref}(t)$.

Let $I_{ref}(\mathbf{z})$ and $I_{input}(\mathbf{z})$ be the reference and input images, respectively, where $\mathbf{z} = (x, y)$. The equation for the illumination environment estimator E_V is

$$E_V = \frac{\sum_{\mathbf{z}\in\mathcal{R}} \| I_{ref}(\mathbf{z}), I_{input}(\mathbf{z}) \|_2}{\sum_{\mathbf{z}\in\mathcal{R}} \| I_{ref}(\mathbf{z}) \|_2}, \tag{4}$$

where $\mathbf{z}$ belongs to the region of interest, $\mathcal{R}$, in the given image. If the modality for face is used, then the region is within the face, and if the modality is for clothes color, then the region corresponds to the body.

For the reference data, noiseless or unbiased data are selected. In many cases, sound and uniformly distributed illumination images are used as the registered signals. In the case of E_A, for example, when the input sound is similarly as clean as the enrolled sound, its value is close to zero, and the nosier the circumstances, the more negative E_A becomes. Similarly, when the illumination of the input image resembles the reference, then E_A approaches zero; otherwise, it approaches a larger positive number.

The relative environment estimator of the face E_F determines, using the threshold δ_F, whether the size of a face is large enough to be identified or whether it is better to choose the clothes color or height. For example, when $E_F > 1$, the face recognition results are highly reliable, and when $E_F << 0.1$ or NaN (that is, the face is not detected), the face becomes less useful.

$$E_F = \frac{W_{in,F} \times H_{in,F}}{W_{ref,F} \times H_{ref,F}}, \tag{5}$$

where $W_{in,F}$ and $H_{in,F}$ are the width and height of the input face, respectively, and $W_{ref,F}$ and $H_{ref,F}$ are the width and height of the reference face, respectively. Note that the face region does not imply the area of a face detection. In a real-world situation, there can be many scenes that have no frontal face; hence, the omega (Ω) shape detection method [17], [19]–[20] is employed, and the upper part of Ω indicates the head, and the lower region of Ω implies the shoulders. Through omega shape tracking, we extract the virtual face region; hence, we can then obtain E_F even when the face of the subject cannot be seen.

3. Stochastic Information Updates

Evidence accumulation [9]–[11] and cue integration [12]–[14] have been attempted to adapt data models to changes in circumstances by updating evidences or cues. However, these approaches do not provide a significant enough impact on person recognition performance because it is difficult to determine which information is good enough to be considered in a model's updating stage.

Let $P(\mathbf{x}_t|M_i = k)$ be the likelihood of the occurrence of input pattern $\mathbf{x}_t$ given that we know the ith modality (M_i) at time t for person $k \in \{1, \ldots, K\}$. We update this likelihood using global and local approaches. In the global approach, $P(\mathbf{x}_t|M_i = k)$ is updated based on the previous likelihood, $P(\mathbf{x}_{t-1}|M_i = k)$. Let $P(\mathbf{x}_t|M_i = k, \alpha)$ be the likelihood within the best α-rank after K likelihoods sorting. Note that there can exist a class in which no member is assigned.

In the local approach, we applied the stochastic likelihood update method using a cumulative binomial statistic, $F(\mathbf{x}|n, p)$. For the n recognition trials, if the rank-1 likelihood corresponds to a correct person with $\omega_i = k$, then we consider it as a success; otherwise, it is considered as a failure, with the probability of success in a single trial denoted by $p = P(\mathbf{x}_t|M_i = k)$.

$$P^{update}(\mathbf{x}_t|M_i = k) = P(\mathbf{x}_{t-1}|M_i = k, \alpha) + \sum_{i=0}^{\alpha} \binom{\alpha}{i} p^i (1-p)^{\alpha-i}. \tag{6}$$

The maximum likelihood of $P^{\text{update}}(\mathbf{x}_t|M_i = k)$, not the original likelihood, determines the recognition performance.

4. Recognizer Set

We use multiple modalities in the proposed framework; face, clothes color, height, and voice. Details of each recognizer are described below.

A. Face Recognition

Face recognition was obtained using AdaBoost training data with modified census transform (MCT) features [21]. An MCT is obtained by calculating the average of the neighborhood intensities and comparing it to the adjacent pixels. AdaBoost trains weak classifiers using both positive and negative samples. To adjust the varying sizes of faces, a cascade method is adopted.

B. Clothes Color Recognition

Clothes color can be a time-limited modality because a person usually wears clothes throughout the day or at least for a certain time period. Three *moments* of color (R, G, B) histograms are applied to recognize clothes color [22]. The first and second moments of a histogram, HM_1 and HM_2, respectively correspond to the mean and standard deviation of the histograms of an image. The third moment of the histogram, HM_3, is evaluated from the following equation:

$$\mathrm{HM}_3 = \left(\frac{1}{b} \sum_{i=1}^{b} \left(i \times \mathrm{H}_i - \bar{\mathrm{H}} \right)^3 \right)^{\frac{1}{3}}, \tag{7}$$

where H_i, b, and $\bar{\mathrm{H}}$ are the ith histogram bin of the image, the number of histogram bins, and the mean of the b histogram bins, respectively.

C. Height Recognition

One's height does not rapidly change on a daily basis; hence, it is considered to be one of the main semi-biometrics. On the other hand, estimation of a person's height is restricted by the fact that the whole of the person's body should appear in an image. We adopt the height calculation method from [22]. A person's height can be calculated using the known position of the camera and the pixel distances of the person's body in the image.

D. Speaker Recognition

Conventional speaker recognition is modeled from a Gaussian mixture model classifier and mel-frequency cepstral coefficient (MFCC) features. We modified this model by adding sub-band likelihood scoring in multiband MFCC [23] and relative autocorrelation sequence (RAS) features.

III. Experiments and Results

We tested our framework on a well-known database, the Extended Yale-B database, and two other databases we collected to reproduce extreme illumination changes and real surroundings.

The second database, of the three, is called "extreme illumination change" (see Fig. 4) and consists of 110 video clips (that is, for every ten persons, one for training and ten for each test). A single video clip has 60 frames. The enrollment data comprises five images per person, where each person is positioned within 1 m of the camera. The test set consists of 60 images per person; there are ten people in total in the test set.

The sample images in Fig. 4 present various illumination effects. The voice corpuses in the second database for multimodal recognition were recorded over twenty sessions using five different distances (1 m, 2 m, 3 m, 4 m, and 5 m) from the microphone and three different types of noise (clean, bubble, and TV). The number of sentences used in the enrollment stage is 10 among 20 (clean type) for the first session, with the speaker being at a distance of 1 m from the camera. The speaker recognition rate, though variable throughout the twenty sessions, declined overall. This is in keeping with the assumption that our voices change over time. The speaker recognizer achieved 79.33% accuracy for 60 sentences (20 sentences × three types of noise) for one out of ten people.

The third database, named "real environment," consists of 25 movie clips per person for 20 people (10 females and 10 males). A movie clip has about 3,200 frames, collected over five sessions. The database consists of five different distances, and the data were collected over a period of twenty days. For those days, the lighting conditions and clothes colors became

Fig. 4. Sample images in an extreme illumination set.

Fig. 5. Sample images of a real environment dataset. A variety of clothe colors and heights are contained in the data. First and third rows are figures with an upper body only (chromatic information is well extracted in these frames). Images in the second and fourth rows feature a whole body, from which the height of the body can be calculated.

Fig. 6. Sample images of various views in uncooperative surroundings. Subjects roamed freely in a 5 m × 5 m room. Subjects' faces are frequently unable to be seen — particularly whenever a person turns left, right, or back. When a person is close to the camera, their face is large enough to obtain a good recognition performance; however, it is then difficult to estimate their height.

Fig. 7. Example images of detected faces in the third dataset. Faces in the first row are taken from images during the enrollment stage; such images have a high E_F value. Images in the second row, which have a low E_F value, have low resolutions and are distorted because they are detected at a long distance. These low-quality images tend to cause worse performance, and E_F can be used to seek out these images.

Fig. 8. Example images of extended Yale-B. First image was captured under direct illumination, and last image was under right-side illumination. First one is a uniformly illuminated image and can therefore be used as a reference image.

increasingly diverse. While the people roamed around the room, as shown in Figs. 5 and 6, we recorded their voices so as to reproduce a real environment. For a movie clip (3,200 frames), each person spoke 30 sentences (one sentence takes about 2 s to 4 s) under one of three different types of noise condition (clean, bubble, and TV sounds). Sensors are located at various places in the test environment because the subject keeps moving around while the camera is recording.

Figures 5 and 6 show sample images from the real environment dataset. The last row of images in Fig. 5 illustrate people whose faces are not detectable but whose heights or color features of clothes are possible to estimate because the whole body can be seen in the scene. For the height, six people in the database belong to the same height group, but the proposed algorithm was able to discriminate them.

The detected faces at a far distance, having low E_F, are easily distorted, as shown in Fig. 7. For example, there are some faces that are difficult to recognize, such as the very small image of the second row, third column and the rotated image in the second row, fifth column, shown in Fig. 7.

1. Results on Extended Yale-B Database

For the evaluation of the environment estimator E_V, we use the Extended Yale-B database [24]–[25], which consists of 38 subjects with varying illumination conditions in the frontal pose. The illumination directions are from '000E+00' to '130E+20,' and '–005+10' to '–130E+20.' We chose 50 out of the 60 illumination types as this was the maximum number of illumination directions that the 38 subjects had in common; the remaining ten illumination types were irregularly present in the database. Figure 8 shows some example images taken from the Extended Yale-B Database. The face with '000E+00' received direct illumination, as shown in the first image of Fig. 8. As a reference image, an unbiased illumination image should be used, for example, the face with '000E+00' in the Extended Yale-B Database.

We calculated E_V for each of the illumination types. In Fig. 9, each subject is represented by a different colored line.

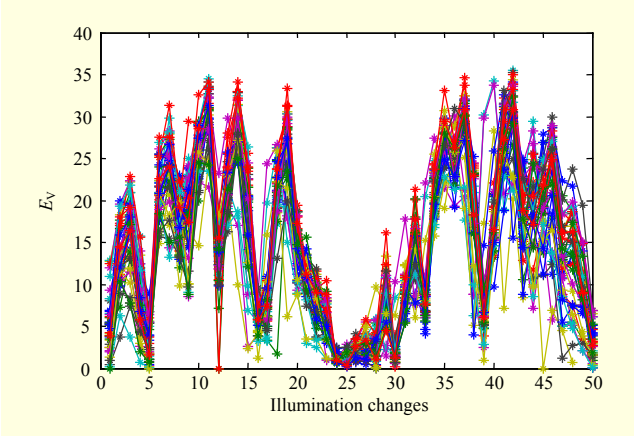


Fig. 9. E_V values of the 38 subjects from the Extended Yale-B database, which contains images of faces with different illumination angles.

Regardless of subjects, the E_V values are quite similar across all changes in illumination. In other words, Fig. 9 shows that E_V depends closely on such illumination changes.

2. Results for Extreme Illumination Change Database

Using the Charf equations, (2) and (3), likelihood m_i is transformed into m_i^{new}. We found that the likelihood normalization enlarged the discriminative characteristic, as shown in Fig. 10. Each bar in Fig. 10 represents a likelihood from a face recognizer taken from the second data set. Figure 10(a) implies that these faces have undistinguishable likelihoods between persons, and after normalization (Fig. 10(b)), the likelihoods of a correctly recognized person are so far apart from those of an incorrectly recognized person.

The recognition performance for the extreme illumination change database is shown in Figs. 11, 12, and 13. The original data, which uses only tracking, presents a considerably lower recognition rate than any other z-based mean update method. We compared the proposed method with the z-based mean update methods by varying z, which is the amount of previous consecutive data, from $z = 3$ to $z = 11$, as shown in Fig. 11.

The recognition accuracy of the z-based mean update methods is degraded at time t compared to $t - 1$, because of the updating of inappropriate information. On the other hand, the proposed method shows a superior performance to all of the z-based mean-update methods featured in Fig. 11. We compared our stochastic method to Lee and others' method [15] with a 5% sorting rule of the discrete probability density function method as shown in Fig. 12, using the extreme illumination change database. Our results are calculated using only the normalization Charf function. The results of Lee and others' method keep fluctuating until the end of the time frame, but our method shows a stable upgrade of the recognition rate during the latter parts of

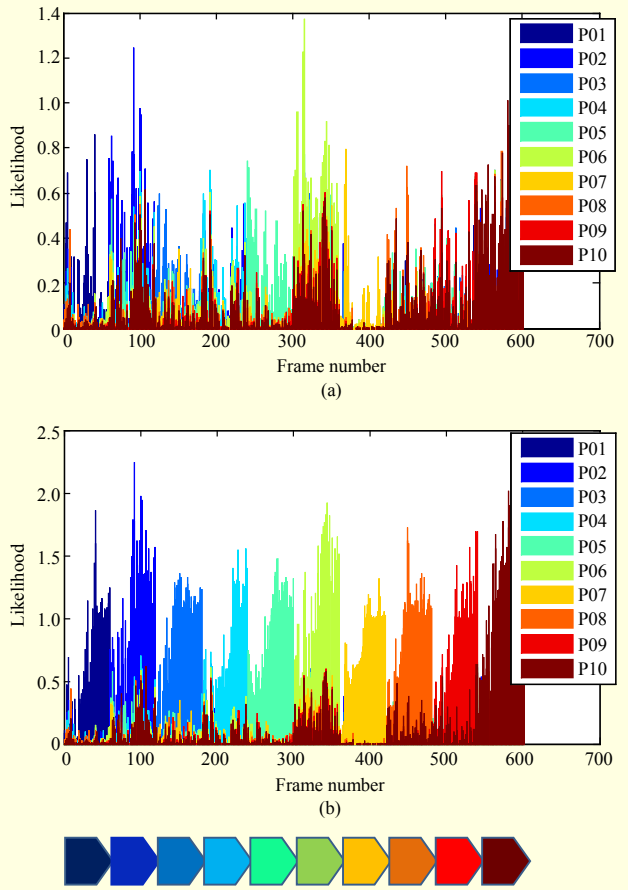


Fig. 10. Normalization and stochastic information updates approach creates a larger difference in likelihood between correct values and other values when applied to the second data set: (a) before normalization and (b) after normalization.

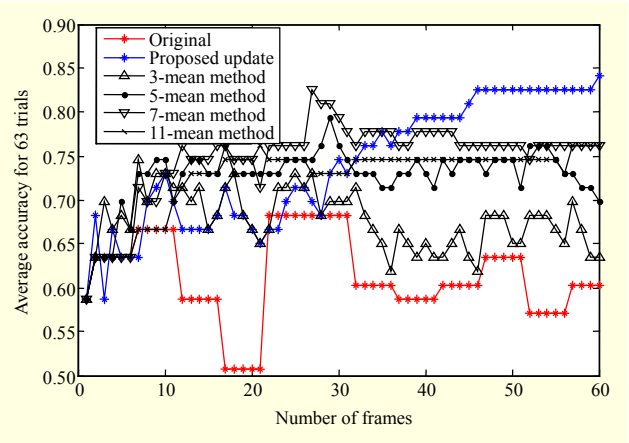


Fig. 11. Recognition performance of the second data set. Original data without z-based mean update method is shown as a red line. Blue line represents the result of the proposed method, and other black lines are z-based mean update methods, where z = 3, 5, 7, and 11 are the lengths of previous reference data.

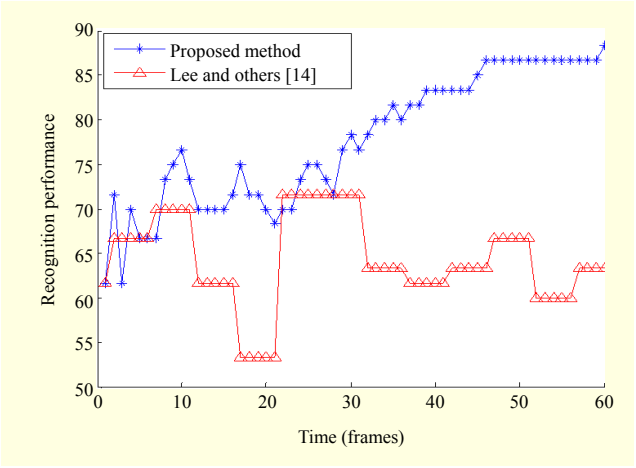


Fig. 12. Comparison of recognition performance. Blue line indicates the results of Lee [15], and red line is the result when applying the method of stochastic information updates.

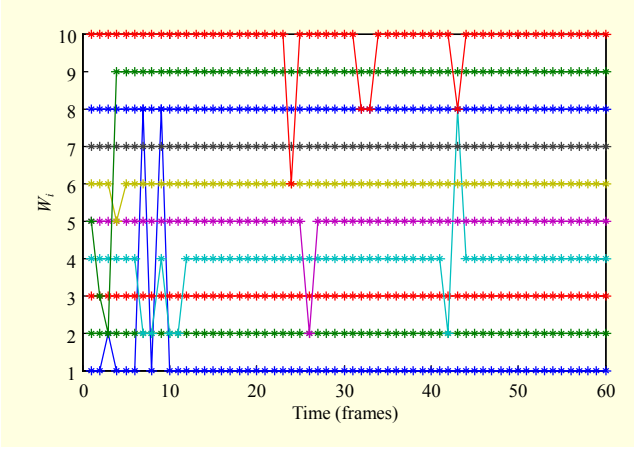


Fig. 13. Evaluation results for the *small* dataset. For ten subjects, this figure shows the recognized recognition ω_i across 60 frames. In the case of ω_i = 1, for example, the lowest blue line, this is confused with ω_i = 8 within 8 frames, but after time passes, the ω_i = 1 test recognizes the correct ω_i.

the same time frame. Figure 13 shows the results of our proposed method (normalization and stochastic information updates) when applied to the extreme illumination change database. In the case of P01 (ω_i = 1) in Table 1, P01 was confused with P02 and P08 in the early part of the given video clip, but P01 was correctly recognized in the latter part of the video clip. Similarly, after the 50 frames had passed, all subjects were correctly recognized.

3. Results on Real Environment Database

Here, a larger database is considered; that is, the real environment database. To reproduce the real world in this database, we let people roam around without any restriction. Therefore, the movie clips in this database contain a wide variation of scale and rotation; illumination changes; and natural views of the human body and face. The voice corpus for this database also contains various types of noise and distance-varying speech.

Table 1. Heights and ω_i's for the third dataset.

ω_i	P01	P02	P03	P04	P05	P06	P07	P08	P09	P10
Height	187	185	185	180	178	177	177	173	169	169
ω_i	P11	P12	P13	P14	P15	P16	P17	P18	P19	P20
Height	164	162	158	158	158	156	149	148	148	146

Table 2. Initial recognition performance of face, height, clothes, and voice (%).

	Face	Height	Clothes color	Voice
Detection	8.62	79.91	84.16	—
Still-to still recognition	49.51	20.86	40.37	42.83
Overall recognition	4.34	19.99	37.32	42.83

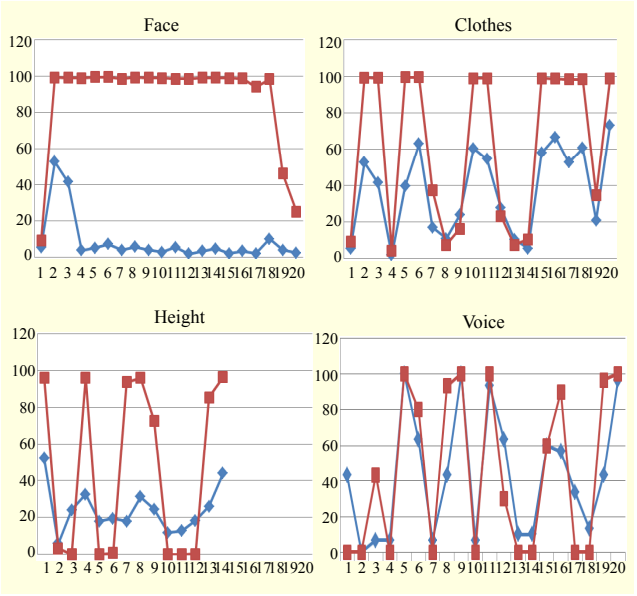


Fig. 14. Recognition performance comparisons before and after applying normalization and stochastic information z-based mean update method for each modality. x-axis indicates the person or group (in the case of height), and the y-axis shows the recognition rate. Blue and red lines represent the performance before and after applying the normalization and stochastic information z-based mean update method, respectively.

ETRI Journal, Volume 37, Number 2, April 2015
http://dx.doi.org/10.4218/etrij.15.0114.0037

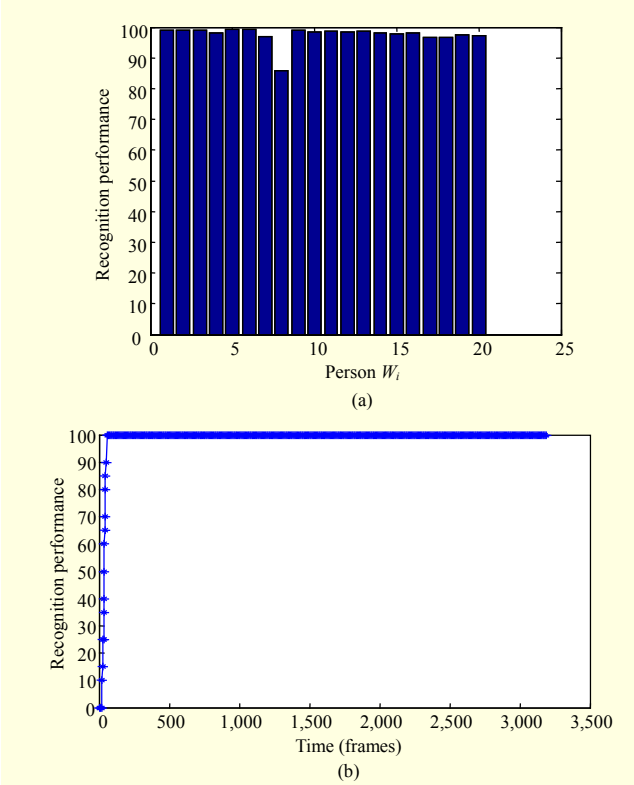


Fig. 15. Recognition performance of the third dataset: (a) applying stochastic information updates and fusion and (b) applying stochastic information updates of fused information based on the environment estimator.

In the case of the height modality, six persons have the same height in the database. This situation can occur in a real environment, too. To deal with this, we made different height groups. The total number of unique height groups is 14 among 20, as is shown in Table 1.

The *initial* performance for the real environment dataset is presented in Table 2. *Initial* implies that we apply only detection, feature extraction, and a recognition step without using either normalization, stochastic information updates, or environment estimators. The face detection rate given in Table 2 is quite low at 8.62%. This is because persons paced back and forth and the frontal face images were barely detected, as shown in Fig. 5. Figure 14 and Table 2 present the initial recognition performance for each modality and the performance for each person, respectively.

In Fig. 14, the original data is illustrated by a blue line, and the proposed method is illustrated by a red line. After applying the method of normalization and stochastic information updates, the performance for each modality is improved for most persons. For the recognition performance of Clothes, in Fig. 14, some people showed a good improvement in terms of recognition performance, while others (ω_i = 1, 4, 7, 8, 12, 13, 14, and 19) showed an improvement over the original

Table 3. Overall recognition rates (%) in the third database.

	Face	Height	Clothes color	Voice	Multi-modal
Recognition rate	92.31	44.71	61.94	44.67	97.59

recognition rate but still at a lower overall performance in general. This is because people may be wearing clothes with similar hue histograms and patterns, as shown in Fig. 5.

Regarding speaker recognition, the number of trials is 30, which differs from the case of a movie clip at 3,200 trials (images were captured at 30 frames per second, but a sentence for a speaker takes 2 s to 3 s). For this reason, the method of normalization and stochastic information updates feebly affects the recognition performance. Therefore, unimodal recognition has constraints and time limitations; thus, unimodal recognition does not achieve a good performance in real situations.

To overcome the unimodal limitation, we adopted the use of environment estimators. These environment estimators played an important role in selecting those modalities that were good enough to identify a person and achieved an overall rate of 97.59%, as shown in Table 3. In the final stage of recognition, the test recorded a rate of 100%, as shown in Fig. 15.

IV. Conclusion

In this paper, we proposed a normalization and stochastic information updates method and environment estimators to improve person recognition performance. The stochastic information updates the likelihood ratios obtained by various modalities. The environment estimators evaluate an input to determine whether it is worthy of being recognized. For the three different databases used in our study, the proposed method recognized uncooperative persons quite well. Our future work will be to try and apply our method to a system that is designed to track and recognize multiple persons.

References

[1] Y.-B. Lee and S. Lee, "Robust Face Detection Based on Knowledge-Directed Specification of Bottom-Up Saliency," *ETRI J.*, vol. 33, no. 4, Aug. 2011, pp. 600–610.

[2] N. Bellotto and H. Hu, "Multisensor Data Fusion for Joint People Tracking and Identification with a Service Robot," *IEEE Int. Conf. Robot Biomimetics*, Sanya, China, Dec. 15–18, 2007, pp. 1494–1499.

[3] D. Jo. et al., "Tracking and Interaction Based on Hybrid Sensing for Virtual Environments," *ETRI J.*, vol. 35, no. 2, Apr. 2013, pp.

356–359.

[4] S. Zhou and R. Chellappa, "Probabilistic Human Recognition from Video," *European Conf. Comput. Vis.*, Copenhagen, Denmark, May 27–31, 2002, pp. 681–697.

[5] S. Zhou and R. Chellappa, "Simultaneous Tracking and Recognition of Human Faces from Video," *IEEE Int. Conf. Acoustic, Speech Signal Process.*, vol. 3, Hong Kong, China, Apr. 6–10, 2003, pp. 225–228.

[6] N. Seo, "*Simultaneous Multi-view Face Tracking and Recognition in Video Using Particle Filtering*," M.S. thesis, Department of Electrical and Computer Engineering, University of Maryland, College Park of Maryland, MD, USA, 2009.

[7] M. Kim et al., "Face Tracking and Recognition with Visual Constraints in Real-World Videos," *IEEE Conf. Comput. Vis. Pattern Recogn.*, Anchorage, AK, USA, June 23–28, 2008, pp. 1–8.

[8] M. Farenzena et al., "Person Re-identification by Symmetry-Driven Accumulation of Local Features," *IEEE Conf. Comput. Vis. Pattern Recogn.*, San Francisco, CA, USA, June 13–18, 2010, pp. 2360–2367.

[9] E.J. Ploran et al., "Evidence Accumulation and the Moment of Recognition: Dissociating Perceptual Recognition Processes Using fMRI," *J. Neurosci.*, vol. 27, no. 44, Oct. 31, 2007, pp. 11912–11924.

[10] W. Kim et al., "Human Action Recognition Using Ordinal Measure of Accumulated Motion," *EURASIP J. Adv. Signal Process.*, Apr. 2010, pp. 1–11.

[11] M. Lucenal et al., "Human Action Recognition Using Optical Flow Accumulated Local Histograms," *IbPRAI*, Povoa de Varzim, Portugal, June 10–12, 2009, pp. 32–39.

[12] M.E. Nilsback and R. Caputo, "Cue Integration through Discriminative Accumulation," *IEEE Conf. Comput. Vis. Pattern Recogn.*, vol. 2, Washington, DC, USA, June 27–July 2, 2004, pp. 578–585.

[13] M.-E. Nilsback, "*A Cue-Integration Scheme for Object Recognition Using Discriminative Accumulation*," M.S. thesis, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden, 2004.

[14] A. Pronobis and B. Caputo, "Confidence-Based Cue Integration for Visual Place Recognition," *IEEE/RSJ Int. Conf. Intell. Robots Syst.*, San Diego, CA, USA, Oct. 29–Nov. 2, 2007, pp. 2394–2401.

[15] J. Lee et al., "Integrating Evidences of Independently Developed Face and Speaker Recognition Systems by Using Discrete Probability Density Function," *IEEE Int. Symp. Robot Human Interactive Commun.*, Jeju, Rep. of Korea, Aug. 26–29, 2007, pp. 667–672.

[16] O. Velek, S. Jaeger, and M. Nakagawa, "*Accumulated-Recognition-Rate Normalization for Combining Multiple On/Off-Line Japanese Character Classifiers Tested on a Large Database*," Multiple Classifier Systems, Guildford, UK: Springer Berlin Heidelberg, vol. 2709, 2003, pp. 196–205.

[17] J. Pelecanos, U. Chaudhari, and G. Ramaswamy, "Compensation of Utterance Length for Speaker Verification," *Proc. ODYSSEY*, Toledo, Spain, May 31–June 3, 2004, pp. 161–164.

[18] H.-J. Kim et al., "Multi-modal User Recognition Based on Environmental Parameters," *Proc. Frontier Comput. Vis.*, Japan, Feb. 2010, pp. 342–345.

[19] M. Li et al., "Rapid and Robust Human Detection and Tracking Based on Omega-Shape Features," *IEEE Int. Conf. Image Process.*, Cairo, Egypt, Nov. 7–10, 2009, pp. 2545–2548.

[20] S. Mukeriee and K. Das, "A Novel Equation Based Classifier for Detecting Human in Images," *Int. J. Comput. Appl.*, vol. 72, no. 6, June 2013, pp. 9–16.

[21] K.-D. Ban et al., "Tiny and Blurred Face Alignment for Long Distance Face Recognition," *ETRI J.*, vol. 33, no. 2, Apr. 2011, pp. 251–258.

[22] D.-H. Kim et al., "A Vision-Based User Authentification System in Robot Environments by Using Semi-Biometrics and Tracking," *IEEE/RSJ Intell. Robots Syst.*, Alberta, Canada, Aug. 2–6, 2005, pp. 1812–1817.

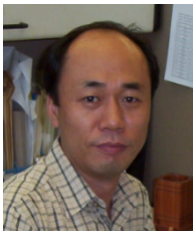
[23] S. Kim, M. Ji, and H. Kim, "Noise-Robust Speaker Recognition Using Subband Likelihoods and Reliable-Feature Selection," *ETRI J.*, vol. 30, no. 1, Feb. 2008, pp. 89–100.

[24] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From Few to Many: Illumination Cone Models for Face Recognition Under Variable Lighting and Pose," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, June 2001, pp. 643–660.

[25] K.-C. Lee, J. Ho, and D.J. Kriegman, "Acquiring Linear Subspaces for Face Recognition Under Variable Lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, May 2005, pp. 684–698.

Hye-Jin Kim received her MS degree in computer science and engineering from Pohang University of Science and Technology, Rep. of Korea, in 2003. She has been a research scientist at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea, since 2005. Her research interests include machine learning, human–robot interaction, and computer vision.

Dohyung Kim received his PhD in computer engineering from Pusan National University, Rep. of Korea, in 2009. He has been a research scientist at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea, since 2002. His research interests are human–robot interaction, computer vision, and pattern recognition.



Jaeyeon Lee received his PhD degree from Tokai University, Japan, in 1996. Since 1986, he has been working as a research scientist at the Electronics and Telecommunications Research Institute, Daejeon, Rep. of Korea. His research interests include robotics, pattern recognition, computer vision, biometric security.

Il-Kwon Jeong received the BS, MS, and PhD degrees in electrical engineering from Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 1992, 1994, and 1999, respectively. Since 1999, he has been with the Department of Visual Content Research, at the Electronics and Telecommunications Research Institute where he was a senior researcher, became a Director in 2008, and a principal researcher in 2010. His current research interests include next generation 3D content, computer graphics, and virtual reality.